Explainable AI for Autonomous Threat Detection in Critical Infrastructure Systems

Edoise Areghan[1] Ogochukwu Susan Ndibe[2]

[1 2]Cybersecurity and Information Assurance, University of Central Missouri, USA

Abstract

Critical infrastructure systems (CIS) including energy, water, transportation, and communication sectors are under continuous cyber-threats from increasingly sophisticated adversaries such as advanced persistent threats (APTs). Artificial intelligence (AI) has enabled autonomous threat detection, yet the opacity of many models poses challenges to trust, transparency, and operational safety. This paper investigates the integration of Explainable Artificial Intelligence (XAI) into AI-driven threat detection for CIS. Using a systematic literature review and thematic analysis of 30 recent studies (2020–2024), the research synthesizes current approaches, frameworks, and implementation challenges. The findings reveal that while XAI improves interpretability, challenges persist regarding trade-offs between accuracy and transparency, adversarial robustness, and human-AI collaboration. A conceptual framework is proposed to guide hybrid XAI adoption, supported by recommendations for human-centric design, policy alignment, and standardization. This work contributes to advancing secure, transparent, and resilient AI systems that reinforce trust and accountability in critical national assets.

Keywords: Explainable AI (XAI); Critical Infrastructure; Cybersecurity; Autonomous Threat Detection; Transparency; Trust; Policy and Resilience.

1 Introduction

Critical infrastructure systems (CIS), encompassing sectors such as energy, water, transportation, and communication, form the bedrock of modern society and national security. The operational continuity and integrity of these systems are under constant threat from increasingly sophisticated cyber-attacks, ranging from advanced persistent threats (APTs) to ransomware campaigns [1]. Traditional security measures, often reliant on static rules and signature-based detection, frequently prove inadequate against rapidly evolving adversarial tactics [2]. This inadequacy propels the development of autonomous threat detection systems, which leverage artificial intelligence (AI) to identify and respond to malicious activities with greater speed and efficacy [3][4].

While AI-driven solutions offer substantial improvements in threat detection capabilities, their "black box" nature can impede human understanding and trust. The opacity of many complex AI models, particularly deep learning architectures, makes it difficult for security analysts to comprehend the reasoning behind a particular alert or classification [5]. This lack of transparency becomes a critical barrier in high-stakes environments like CIS, where misinterpretations can lead to catastrophic operational failures, economic disruption, or even loss of life. Consequently, the field of Explainable AI (XAI) has

emerged to address this challenge, seeking to render AI decisions intelligible to human operators [6].

The integration of XAI into autonomous threat detection for CIS is not merely a technical advancement but a strategic imperative. It facilitates human oversight, fosters trust, and enables more informed decision-making during critical incidents. Explanations from AI systems can help security personnel understand attack vectors, refine detection models, and develop more robust defense strategies. Without clear explanations, the adoption and effective utilization of AI in such sensitive contexts may be hindered, limiting its potential to enhance overall cybersecurity posture and resilience [7].

This paper offers an in-depth examination of the application of Explainable AI for autonomous threat detection within critical infrastructure systems. It synthesizes current research, identifies key challenges, and proposes integration strategies. The subsequent sections will detail the methodological approach, review relevant literature concerning critical infrastructure threats, AI advancements in detection, and the principles of XAI. Following this, an analysis and discussion will address implementation challenges, integration frameworks, policy implications, and future research opportunities. The overarching objective is to contribute to a comprehensive understanding of how XAI can bolster the security and trustworthiness of AI-driven defenses in vital national assets.

2 Methodology

A systematic literature review and thematic analysis approach guided this research. The initial phase involved an extensive search across prominent academic databases, including IEEE Xplore, ACM Digital Library, Scopus, and Web of Science. Search terms focused on combinations of "Explainable AI," "XAI," "interpretable AI," "autonomous threat detection," "cybersecurity," "critical infrastructure," "industrial control systems," "SCADA," "APT," and "resilience." The search period was primarily constrained to publications from 2020 to 2024 to capture the most recent advancements and perspectives in this rapidly evolving field.

Inclusion criteria for selected articles prioritized peer-reviewed journal articles, conference papers, and comprehensive review papers that directly addressed the intersection of AI-driven threat detection, explainability, and critical infrastructure security. Papers focusing solely on general AI applications or XAI in unrelated domains were excluded. Additionally, publications lacking empirical validation or robust theoretical frameworks were deselected. The initial search yielded several hundred results, which were then filtered based on title, abstract, and keyword relevance. A total of 30 core articles were identified for in-depth analysis, supplemented by additional relevant foundational works as necessary.

The selected literature underwent a rigorous thematic analysis. This process involved several iterative steps: initial coding, theme identification, theme refinement, and categorization. During initial coding, key concepts, methodologies, findings, and challenges from each paper were extracted and annotated. Examples of codes included "AI model opacity," "SHAP values," "adversarial attacks," "false positives," "operator trust," and "policy implications." These codes were then grouped into broader themes

such as "critical infrastructure vulnerabilities," "AI detection techniques," "XAI methods," "human-AI collaboration," "regulatory concerns," and "resilience strategies."

Thematic analysis facilitated the identification of gaps in current research, convergent and divergent perspectives across studies, and emerging trends. Particular attention was given to articles that provided concrete examples of XAI implementation in cybersecurity contexts, discussed the practical benefits and drawbacks of different XAI techniques, or highlighted the unique requirements of critical infrastructure protection. The synthesis of this thematic analysis forms the basis for the literature review, providing a structured overview of the relevant knowledge landscape. Furthermore, this methodological framework informed the subsequent analysis and discussion sections, allowing for a comprehensive evaluation of current challenges, potential solutions, and future research trajectories in the context of XAI for autonomous threat detection in critical infrastructure systems.

3 Literature Review and Thematic Analysis

3.1 Critical Infrastructure Threat Landscape

Critical National Infrastructure (CNI) faces a sophisticated and ever-present threat landscape, characterized by actors ranging from state-sponsored entities to cybercriminals. These essential assets, including energy grids, water systems, transportation networks, and communication frameworks, are increasingly digitized and interconnected, expanding their attack surfaces [8]. The consequences of successful attacks against CNI extend beyond financial losses, potentially leading to widespread societal disruption, public safety hazards, and national security compromises.

Advanced Persistent Threats (APTs) represent a particularly insidious category of cyber-attack, characterized by their sophisticated orchestration, stealthy execution, extended persistence, and targeted nature [1][9]. These threats often evade traditional intrusion detection systems, which rely on static rules or predefined signatures [2]. Attackers continually adapt their methods, employing polymorphic malware, zero-day exploits, and social engineering to bypass defenses and maintain covert access within targeted systems [4]. For instance, the compromise of trust and reputation through social media campaigns represents a potential attack vector against critical information infrastructures that were formerly public utilities, impacting public perception and stability [10].

The sheer volume and velocity of data generated within CNI environments make manual threat detection impractical and often ineffective. Security teams are overwhelmed by alerts, leading to alert fatigue and potentially missing critical indicators of compromise [4]. This creates an urgent requirement for autonomous detection systems that can process vast datasets, identify anomalies, and flag potential threats in real-time with minimal human intervention. Such systems must possess the capability to identify novel and evasive attack patterns, a task where traditional methods frequently fall short.

The shift from risk analysis to resilience as a dominant paradigm underscores the recognition that complete prevention of adverse events, such as cyber-attacks, is often unattainable [11]. Instead, the focus has expanded to designing systems that can

withstand and rapidly recover from disruptive events. Autonomous threat detection, particularly when enhanced by AI, directly supports this resilience objective by enabling faster identification and containment of incidents, thereby reducing their overall impact and facilitating quicker restoration of normal operations [7].

3.2 AI-Driven Detection Technique in Critical Infrastructure

The application of Artificial Intelligence (AI) has significantly transformed the landscape of threat detection in critical infrastructure. AI models offer the capacity to process and analyze immense volumes of data, identifying subtle patterns and anomalies indicative of malicious activity that human analysts or rule-based systems might overlook [8][4]. Various AI techniques have found utility in this domain, including machine learning (ML), deep learning (DL), and reinforcement learning (RL) [7].

Machine learning algorithms, such as Support Vector Machines (SVMs), Logistic Regression, Random Forests, and Gradient Boosting Machines, are employed for tasks like malware detection, intrusion detection, and anomaly detection. For instance, in detecting malicious PDF files, the Lightweight Gradient Boosting Machine (LGBM) has shown high accuracy, precision, and F1-scores, demonstrating the efficacy of ML in identifying evasive threats [5]. Similarly, credit card fraud detection benefits from ML techniques, with models achieving high accuracy rates, though recall metrics often provide a more comprehensive comparison [12].

Deep learning models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks, excel at learning complex representations from raw data, making them suitable for detecting sophisticated and novel cyber threats. Their ability to automatically extract features from diverse data types, such as network traffic logs, system calls, and executable binaries, reduces the need for manual feature engineering [7]. For example, a novel deep learning-based threat-hunting model combining Parallel Stacked Long Short-Term Memory (PSLSTM) networks with a multi-head attention mechanism has been proposed for improved attack detection in smart healthcare systems [13].

The integration of AI extends to addressing advanced persistent threats (APTs), which often leverage sophisticated, long-term, and stealthy network intrusions [1]. Approaches like provenance graph-based kernel-level auditing, though promising, encounter difficulties in reconstructing complex lateral attack chains and detecting dynamic evasion behaviors [9]. AI can augment these methods by identifying patterns of behavior that signify an APT, even when individual actions appear innocuous. Threat Insight, for example, uses threat data mining, feature modeling, and semantic reasoning based on an APT Threat Intelligence Knowledge Graph to detect early-stage threats by analyzing IPs from Honey Points, minimizing reliance on retrospective audit logs [2].

Furthermore, Generative AI and Large Language Models (LLMs) are beginning to be explored for enhancing Critical Infrastructure Protection (CIP), offering capabilities in threat intelligence generation, vulnerability assessment, and even automated response generation. Agentic AI for proactive defense mechanisms is also under consideration, pushing the boundaries of autonomous security operations. These advancements collectively demonstrate AI's capacity to bolster real-time threat detection and counter

cyber threats in critical infrastructure, as seen in cases like the Colonial Pipeline ransomware attack, highlighting both AI's capabilities and limitations [3].

3.3 Explainability, Transparency, and Trust

The increasing reliance on AI for autonomous threat detection, particularly in sensitive environments like critical infrastructure, amplifies the importance of explainability, transparency, and trust. While AI models can achieve high accuracy in detecting threats, their complex internal workings often render them opaque "black boxes" [5]. This opacity presents significant hurdles for security analysts and operators who require a clear understanding of why a specific alert was triggered or why a particular decision was made [13].

Explainability, often achieved through Explainable AI (XAI) techniques, seeks to make AI models more understandable to humans. The goal is not merely to provide accurate predictions but to offer insights into the reasoning process, thereby enabling human users to comprehend, trust, and effectively manage AI systems. For example, the use of Shapley Additive Explanations (SHAP) values can identify crucial features influencing a model's prediction, enhancing interpretability and even model performance [5]. Without such explanations, operators might be hesitant to fully trust or act upon AI-generated alerts, especially when false positives or false negatives could have severe consequences in a critical infrastructure context.

Transparency, a closely related concept, refers to the degree to which an AI system's operation is understandable. It encompasses various dimensions, including data transparency (understanding the input data), model transparency (understanding the internal mechanisms), and decision transparency (understanding why a particular output was produced) [6]. In security systems, transparency is essential for debugging, auditing, and validating the AI's behavior. If an AI system behaves unexpectedly or makes an erroneous decision, a lack of transparency hinders the ability to diagnose the root cause and implement corrective measures.

Trust in AI systems is paramount for their successful deployment in critical infrastructure. Operators must trust that the AI is reliable, fair, and secure. XAI contributes to building this trust by demystifying the AI's decisions, allowing humans to verify the logic and identify potential biases or vulnerabilities [13]. However, it is also acknowledged that transparency can, in certain contexts, cause harm, particularly concerning privacy or multi-agent game theory scenarios where revealing too much information might be exploited by adversaries. Therefore, a balance must be struck between providing sufficient explanation for human understanding and maintaining security through partial opacity where strategically necessary. The concept of zero-trust architecture (ZTA) in AI model deployment within cloud environments emphasizes skepticism even towards trusted entities, highlighting the inherent challenges in establishing trust in complex AI systems [14].

The policy landscape is also responding to the necessity for explainability. Governmental policies are attempting to establish regulatory baselines for AI explainability yet often struggle with a consensus on what constitutes a valid algorithmic explanation and how to ensure its practical implementation and usability across diverse stakeholders [6]. This

ongoing development underscores the critical relationship between technical advancements in XAI and the broader socio-technical and regulatory frameworks required for secure and trustworthy AI deployment in critical sectors.

3.4 Frameworks and Methods for Explainable AI

The field of Explainable AI (XAI) offers various frameworks and methods to enhance the interpretability of AI models used in threat detection, addressing the "black box" problem. These methods can broadly be categorized into intrinsic interpretability and post-hoc explainability techniques. Intrinsic interpretability involves using models that are inherently understandable, such as decision trees or linear models, while post-hoc methods apply explanation techniques to complex, pre-trained models.

One prominent post-hoc technique is Shapley Additive Explanations (SHAP), which provides local explanations by attributing the contribution of each feature to a specific prediction [5]. SHAP values, derived from cooperative game theory, quantify the impact of each feature by considering all possible combinations of features, thereby offering a robust and theoretically sound measure of feature importance. For example, in malicious PDF detection, SHAP values guided feature engineering, resulting in an improved classification model by identifying crucial features [5]. This local interpretability allows security analysts to understand precisely why a particular file was flagged as malicious, pointing to specific characteristics within the file.

Another widely used method is Local Interpretable Model-agnostic Explanations (LIME). LIME works by approximating the behavior of a black-box model with a simpler, interpretable model (e.g., a linear model) around the vicinity of a specific prediction. This local fidelity helps explain individual predictions by highlighting the features that are most influential for that instance. While SHAP offers a more rigorous theoretical foundation, LIME provides quick, intuitive explanations, which can be valuable for real-time threat analysis.

Global interpretability methods complement local explanations by providing an overall understanding of the model's behavior across its entire dataset. Techniques such as permutation feature importance, partial dependence plots (PDPs), and individual conditional expectation (ICE) plots help reveal how features generally influence model predictions. For instance, a global surrogate model, like an interpretable decision tree, can be trained to mimic the behavior of a complex black-box model, offering a simplified, overarching view of its decision logic [5]. This can assist in identifying systemic biases or unexpected correlations learned by the AI.

The integration of XAI with blockchain technology represents an innovative approach to enhance trust and transparency in AI-driven threat detection. By using blockchain to validate and store data between multiple cloud vendors, it can mitigate risks associated with malicious cloud providers supplying false information, thereby safeguarding the integrity of data used for threat hunting [13]. This combination fosters a more secure and accountable environment for AI deployment in sensitive sectors like smart healthcare systems, which share characteristics with critical infrastructure in terms of their vulnerability and impact.

Furthermore, XAI frameworks are evolving to incorporate user-centric design principles, recognizing that the effectiveness of an explanation depends on the needs and expertise of the end-user. This includes tailoring explanations to different stakeholders, such as incident responders, system administrators, or regulatory bodies. For example, a security analyst might require detailed feature importance, while a policy maker might need a high-level overview of the AI's risk assessment strategy. These methods collectively empower human operators to better understand, evaluate, and ultimately trust the autonomous decisions made by AI systems in critical infrastructure security.

Figure 1. Conceptual Framework for Explainable Autonomous Threat Detection in Critical Infrastructure

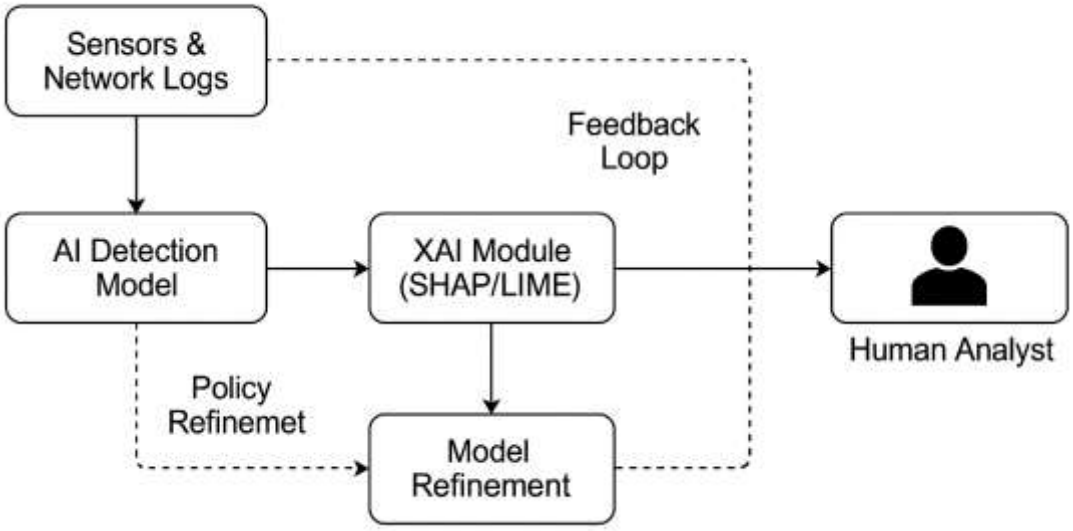

Figure 1 illustrates the conceptual workflow of explainable autonomous threat detection within critical infrastructure systems. Data from sensors and network logs are processed by AI-based detection models, whose decisions are interpreted through an XAI module employing methods such as SHAP or LIME. Explanations are presented to human analysts, who validate or refine these outcomes and feed insights back into model and policy updates. This continuous feedback loop ensures adaptability, transparency, and improved decision-making, strengthening both the technical and human dimensions of cyber resilience.

Table 1. Comparison of Major XAI Techniques for Threat Detection

Technique	Type	Interpretability	Advantages	Limitations	Representative Use Case
SHAP	Post-hoc	Local/Global	Theoretically rigorous, consistent	Computationally heavy	Malicious PDF detection (Al-Fayoumi et al., 2024)
LIME	Post-hoc	Local	Fast, intuitive explanations	Approximation errors	Intrusion alerts in ICS
Decision Trees	Intrinsic	Global	Simple visual logic	Low scalability	Rule-based anomaly detection
Counterfactuals	Post-hoc	Local	Human-friendly "what-if" reasoning	Hard to generate for high-dim. data	Attack attribution
Surrogate Models	Post-hoc	Global	Summarize black-box logic	May misrepresent model	Policy audit visualization

Table 1 compares prominent explainable AI techniques such as SHAP, LIME, decision-tree models, counterfactual reasoning, and surrogate modeling highlighting their interpretability scope, advantages, limitations, and representative cybersecurity use cases. The comparison emphasizes how each technique contributes uniquely to the transparency accuracy balance required in critical infrastructure defense, guiding researchers and practitioners toward selecting context-appropriate XAI methods

4 Analysis and Discussion

4.1 Challenges in Implementing Explainable AI for Autonomous Threat Detection

Implementing Explainable AI (XAI) in autonomous threat detection for critical infrastructure systems presents several significant challenges, stemming from both technical complexities and the unique operational requirements of these sensitive environments. One primary challenge involves the inherent trade-off between model complexity, accuracy, and interpretability. Highly accurate AI models, especially deep learning architectures, often achieve their performance through intricate, non-linear computations that are difficult to unravel. Simplifying these models for explainability can

sometimes compromise their detection accuracy, which is unacceptable in scenarios where even a slight reduction in performance could lead to missed threats or catastrophic failures.

The dynamic and adversarial nature of cyber threats introduces another layer of complexity. Attackers constantly evolve their techniques, making threat detection a continuous arms race [4]. XAI explanations, if not carefully designed, could potentially reveal vulnerabilities in the AI model itself, providing adversaries with insights to craft more effective evasion strategies. This necessitates a careful balance between providing sufficient transparency for human operators and maintaining strategic opacity to prevent exploitation. The concept of transparency causing harm in certain contexts, particularly where privacy or adversarial game theory is involved, underscores this delicate balance.

Data quality and availability within critical infrastructure environments also pose hurdles for XAI implementation. Training robust AI models, let alone explainable ones, requires vast amounts of high-quality, labeled data, including normal operational data and various attack scenarios. Such data can be scarce, proprietary, or difficult to collect due to operational sensitivities and privacy concerns [15]. Explanations generated from biased or incomplete datasets may themselves be misleading, eroding trust rather than building it. Furthermore, the operational speed required for autonomous threat detection can conflict with the computational overhead often associated with generating detailed explanations, which may delay critical response actions.

Integrating XAI tools into existing critical infrastructure security operations centers (SOCs) also poses practical challenges. Security analysts require explanations that are actionable, intuitive, and seamlessly integrated into their workflows, without adding unnecessary cognitive load or requiring extensive retraining. The current disparity in XAI tools and the lack of standardized metrics for evaluating explanation quality further complicate this integration [6]. Finally, the regulatory and policy landscape for AI explainability is still nascent, lacking clear guidance and consensus on what constitutes adequate explanation, making compliance difficult for organizations operating critical infrastructure [6]. These multifaceted challenges demand innovative technical solutions, robust testing frameworks, and careful consideration of human-AI interaction principles.

4.2 Integration Strategies for XAI in Critical Infrastructure Systems

Effective integration of Explainable AI (XAI) into critical infrastructure systems (CIS) requires a multi-faceted strategy that addresses technical, operational, and human-centric considerations. A fundamental approach involves deploying XAI techniques incrementally, starting with less critical functions or in an offline analysis capacity, before full operational integration. This phased adoption allows for thorough testing, validation, and user familiarization without immediate high-stakes consequences.

1. **Hybrid XAI Architectures:** Combining intrinsically interpretable models with post-hoc explanation techniques can leverage the strengths of both. Simple, transparent models (e.g., decision trees, rule-based systems) can handle routine, well-understood threats, providing clear explanations. For complex, novel threats requiring sophisticated black-box models (e.g., deep neural networks), post-hoc methods like SHAP or LIME can generate local explanations for specific

detections [5]. This hybrid approach ensures a baseline of interpretability while maintaining high detection accuracy.

2. **Contextual Explanations and User-Centric Design:** Explanations must be tailored to the specific context and the user's role and expertise. Security analysts require detailed technical insights (e.g., feature importance, anomaly scores), while incident commanders might need high-level summaries of potential impact and recommended actions. Developing dashboards that visualize XAI outputs intuitively, using natural language generation for explanation narratives, and providing interactive tools for exploring model reasoning can enhance usability.
3. **Human-in-the-Loop Validation:** Integrating XAI within a human-in-the-loop framework is crucial. This involves human operators reviewing AI-generated explanations, providing feedback, and validating or overriding autonomous decisions. This iterative process refines both the AI model and its explanation capabilities, building trust and improving overall system performance. Blockchain-enabled XAI, as proposed for smart healthcare systems, offers a mechanism to validate and store data, thereby enhancing the trustworthiness of AI decisions in a distributed environment [13].
4. **Adversarial Robustness for XAI:** Given the adversarial nature of cyber threats, XAI methods themselves need to be robust against manipulation. Research into "adversarial XAI" is vital to ensure that explanations cannot be easily faked or exploited by attackers to mask their activities. This involves designing explanation techniques that are resilient to minor perturbations in input data that could otherwise lead to misleading explanations.
5. **Standardization and Benchmarking:** The development of industry standards and benchmarks for evaluating XAI effectiveness in cybersecurity contexts is essential. Clear metrics for explanation quality, fidelity, and utility will accelerate adoption and facilitate the comparison of different XAI techniques. This also supports the evolving regulatory landscape which currently lacks a consensus on what constitutes a valid algorithmic explanation [6].

By systematically applying these integration strategies, critical infrastructure operators can harness the power of autonomous AI threat detection while ensuring that human oversight, understanding, and trust remain central to their security posture.

4.3 Human Factors and Cognitive Considerations

Effective deployment of XAI depends not only on algorithms but on human cognition and trust dynamics. Security analysts must interpret explanations rapidly during high-pressure incidents. Poorly designed interfaces or verbose explanations can increase cognitive load, delaying responses. Research in Human-Computer Interaction (HCI) emphasizes concise visual summaries and adjustable explanation depth depending on expertise.

Figure 2. Example Explanation in the Domain of Critical Infrastructure

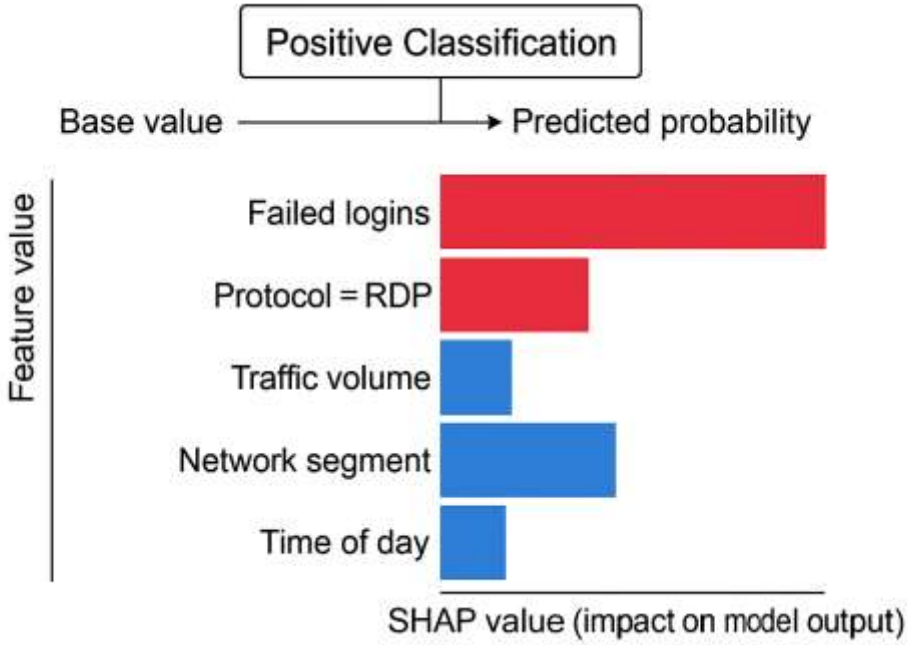

Figure 2 provides a domain-specific illustration of model explainability using SHA values. It visualizes how individual input features such as failed logins, RDP protocol use, network segment, traffic volume, and time of day influence the AI system's prediction of a potential threat. Red bars denote positive contributions toward malicious classification, while blue bars indicate mitigating factors. Such localized visual explanations assist analysts in understanding why alerts were triggered, enabling faster validation, reducing cognitive load, and enhancing overall trust in autonomous detection systems

Trust calibration neither over-trusting nor under-trusting AI is essential. Feedback mechanisms where analysts rate explanation usefulness can refine model presentation. Understanding these psychological and ergonomic elements ensures XAI enhances rather than hinders decision quality.

4.4 Policy and Governance Considerations

Regulators increasingly demand algorithmic transparency in safety-critical sectors. However, definitions of "sufficient explanation" differ across jurisdictions (EU AI Act, US NIST AI RMF, UK DSIT Principles). For CIS operators, compliance must harmonize with operational secrecy and national-security requirements.

Policies should:

1. Mandate explainability thresholds for autonomous cyber-defense systems.
2. Require audit trails for AI decisions.
3. Support data-sharing frameworks for validating XAI models under controlled conditions.

4. Incentivize interdisciplinary collaboration between technologists, ethicists, and legal experts.

This alignment strengthens public trust and ensures accountability in automated defense infrastructures.

Figure 3. Policy–Technology–Resilience Integration Model

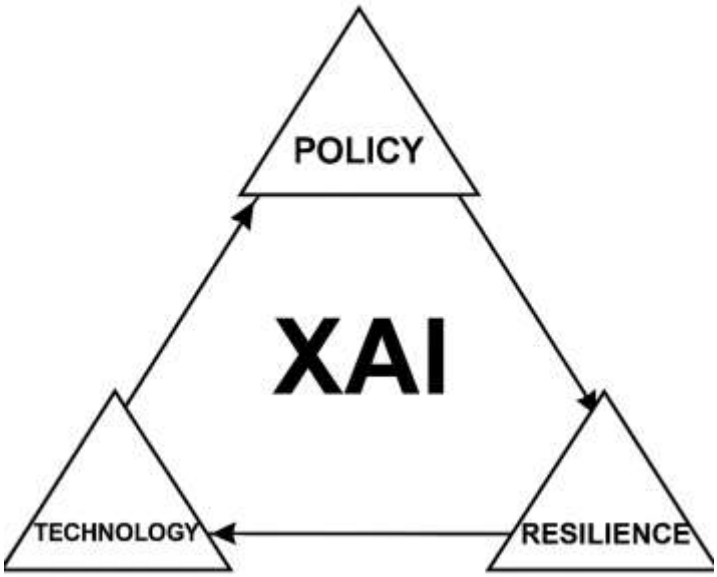

Figure 3 presents the Policy–Technology–Resilience Integration Model, demonstrating how XAI acts as a central linking mechanism among these three domains. The technological layer comprises AI and XAI algorithms driving threat detection; the policy layer provides regulatory oversight, standards, and accountability; and the resilience layer encompasses human oversight, adaptive recovery, and learning. Arrows indicate feedback among the three layers, showing that explainable AI both informs and is constrained by evolving policy frameworks while simultaneously reinforcing system resilience through transparent decision pathways.

4.5 Implications for Security, Resilience, and Policy

The integration of Explainable AI (XAI) into autonomous threat detection for critical infrastructure systems (CIS) carries profound implications across security, resilience, and policy domains. From a security perspective, XAI can significantly enhance the effectiveness of threat detection and response. By demystifying AI decisions, security analysts gain a deeper understanding of attack vectors and indicators of compromise, enabling them to refine detection rules, harden vulnerabilities, and develop more targeted mitigation strategies. This improved comprehension reduces false positives and false negatives, which are costly in CIS, preventing alert fatigue and ensuring that critical threats receive prompt attention. The ability to understand why a system flagged a particular activity as malicious can also aid in post-incident analysis, helping reconstruct attack paths and attributing threats more accurately [2].

For resilience, XAI contributes by fostering trust and improving human-AI collaboration during disruptive events. Resilience, defined as the capacity to withstand and recover from adverse events, relies heavily on timely and accurate decision-making [16]. When human operators trust the AI's explanations, they are more likely to accept its recommendations and act decisively during a cyber-attack. This trust reduces friction and hesitation, thereby accelerating response times and minimizing the impact of incidents. XAI also facilitates the continuous learning and adaptation of security systems, as human feedback based on understandable explanations can be used to improve AI models, making them more robust against evolving threats [4]. This contributes to the overall cyber resilience of critical infrastructure by enhancing its adaptive capacity.

On the policy front, XAI addresses growing concerns regarding accountability, ethical deployment, and regulatory compliance of AI in high-stakes applications. As AI systems assume greater autonomy in protecting CIS, there is an increasing demand for mechanisms to ensure they operate responsibly and predictably [6]. Policy frameworks need to evolve to mandate explainability requirements, particularly for systems making critical decisions that could impact public safety or national security. This includes defining what constitutes an adequate explanation, establishing standards for auditing AI decisions, and clarifying liability in cases of AI-induced failures. The current lack of a common regulatory baseline for XAI makes it difficult for organizations to ensure compliance and for policymakers to enforce responsible AI development [6]. Therefore, policies must encourage research into robust XAI techniques, facilitate data sharing for model validation, and promote interdisciplinary collaboration between AI developers, cybersecurity experts, and legal scholars to forge comprehensive guidelines for secure and trustworthy AI deployment in critical infrastructure. The application of hazard analysis approaches, such as the Taxonomy for AI Hazard Analysis (TAIHA), also becomes relevant here, extending traditional models to account for AI as a causal factor in accidents within safety-critical systems [17].

5 Future Research Directions and Opportunities

The convergence of Explainable AI (XAI) and autonomous threat detection in critical infrastructure systems (CIS) opens numerous avenues for future research and innovation. Addressing the current limitations and enhancing the practical applicability of XAI in this domain will require concerted effort across several key areas.

1. **Real-time Explainability and Performance Optimization:** Current XAI methods can incur significant computational overhead, which conflicts with the real-time demands of autonomous threat detection in CIS. Future research should focus on developing lightweight, efficient XAI algorithms capable of generating explanations instantaneously without compromising detection speed or accuracy. This could involve exploring hardware-accelerated XAI, optimizing existing algorithms, or devising novel explanation techniques that are intrinsically fast.
2. **Adversarial XAI and Robustness:** Given the sophisticated nature of cyber threats, future work needs to investigate the adversarial robustness of XAI methods. Attackers may attempt to manipulate AI models to generate misleading explanations or hide their malicious activities within seemingly benign outputs.

Research efforts should concentrate on designing XAI techniques that are resilient to such adversarial attacks and on methods for detecting when explanations themselves might be compromised. This includes exploring techniques to balance transparency with strategic opacity to avoid providing adversaries with exploitable information.

3. **Multi-modal and Context-Aware Explanations:** Critical infrastructure systems generate diverse data types, including network traffic, sensor readings, control commands, and human operator logs. Future XAI research should focus on developing methods that can provide coherent explanations by integrating information from these multi-modal sources. Furthermore, explanations should be context-aware, adapting to the specific operational state of the CIS, the type of threat, and the expertise level of the human operator receiving the explanation.
4. **Standardization, Evaluation, and Benchmarking:** The field of XAI for cybersecurity lacks widely accepted standards and benchmarks for evaluating the quality and utility of explanations. Future research must contribute to establishing objective metrics for assessing clarity, fidelity, stability, and actionability of XAI outputs in CIS contexts. Developing standardized datasets that incorporate ground truth explanations for various cyber-attacks would greatly facilitate comparative studies and advancement in the field.
5. **Human Factors and Cognitive Load:** Understanding how human operators perceive, interpret, and act upon AI explanations is crucial. Future research should involve interdisciplinary studies combining AI, cybersecurity, and human-computer interaction (HCI) to optimize the presentation of explanations, minimize cognitive load, and maximize trust and effectiveness. This includes exploring the psychological aspects of trust in AI and developing user interfaces that effectively communicate complex AI reasoning.
6. **Regulatory and Ethical Frameworks:** Continued collaboration between technologists, policymakers, and legal experts is needed to develop comprehensive regulatory and ethical frameworks for XAI in CIS. Research can inform these frameworks by exploring the feasibility and impact of different explainability requirements, accountability mechanisms, and privacy considerations. This includes addressing the challenges highlighted by current policy analyses regarding the definition, feasibility, and usability of explanations [6].
7. **Proactive XAI for Threat Hunting and Prediction:** Beyond reactive threat detection, XAI could be leveraged for proactive threat hunting and predictive analytics. Explanations could help security analysts understand why certain system behaviors might indicate pre-attack reconnaissance or early-stage compromise, allowing for preemptive action. This also connects to the advancements in LLMs and Agentic AI for proactive defense mechanisms in CIP.

By addressing these areas, the scientific community can significantly advance the deployment of trustworthy, effective, and resilient AI-driven security solutions for critical infrastructure.

6 Conclusion

6.1 Summary

Autonomous AI has revolutionized threat detection for critical infrastructures, yet its opacity introduces operational risk. XAI bridges this divide by enabling transparent reasoning and human oversight. Through literature synthesis, this paper demonstrates that hybrid XAI frameworks combining interpretable and opaque models can deliver both accuracy and accountability.

6.2 Contributions and Practical Implications

1. Scholarly Contribution: Provides a structured synthesis of XAI methods contextualized to CIS security.
2. Practical Framework: Proposes a conceptual model integrating detection, explanation, and human validation.
3. Policy Insight: Identifies governance actions to standardize explainability and trust metrics.
4. Future-Facing Direction: Outlines R&D priorities real-time explainability, adversarial robustness, and cognitive optimization.

These contributions position XAI not merely as a transparency tool but as a strategic enabler of national cyber-resilience, ensuring that AI-driven security remains both autonomous and accountable.

References

[1] B. Zhang, Y. Gao, B. Kuang, C. Yu, A. Fu, and W. Susilo, "A Survey on Advanced Persistent Threat Detection: A Unified Framework, Challenges, and Countermeasures," *ACM Computing Surveys*, vol. 57, no. 3. Association for Computing Machinery (ACM), pp. 1–36, Nov. 11, 2024. doi: 10.1145/3700749.

[2] Z. Wang, Y. Zhou, H. Liu, J. Qiu, B. Fang, and Z. Tian, "ThreatInsight: Innovating Early Threat Detection Through Threat-Intelligence-Driven Analysis and Attribution," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 12. Institute of Electrical and Electronics Engineers (IEEE), pp. 9388–9402, Dec. 2024. doi: 10.1109/tkde.2024.3474792.

[3] Bright Ojo and Chukwudi Tabitha Aghaunor, "AI-driven cybersecurity solutions for real-time threat detection in critical infrastructure," *International Journal of Science and Research Archive*, vol. 12, no. 2. GSC Online Press, pp. 1716–1726, Aug. 30, 2024. doi: 10.30574/ijsra.2024.12.2.1401.

[4] S. Verma, P. Pali, P. Kori, and M. Tiwari, "Advanced Threat Detection Systems for Protecting Critical Infrastructure," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 11, no. 06. Ess & Ess Research Publications, pp. 8896–8901, Jun. 24, 2023. doi: 10.15680/ijircce.2023.1106077.

[5] M. Al-Fayoumi, Q. Abu Al-Haija, R. Armoush, and C. Amareen, "XAI-PDF: A Robust Framework for Malicious PDF Detection Leveraging SHAP-Based Feature Engineering," *The International Arab Journal of Information Technology*, vol. 21, no. 1. Zarqa University, Jan. 01, 2024. doi: 10.34028/iajit/21/1/12.

[6] L. Nannini, A. Balayn, and A. L. Smith, "Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK," *2023 ACM Conference on Fairness Accountability and Transparency*. ACM, pp. 1198–1212, Jun. 12, 2023. doi: 10.1145/3593013.3594074.

[7] M. Al-Hawawreh, Z. Baig, and S. Zeadally, "AI for Critical Infrastructure Security: Concepts, Challenges, and Future Directions," *IEEE Internet of Things Magazine*, vol. 7, no. 4. Institute of Electrical and Electronics Engineers (IEEE), pp. 136–142, Jul. 2024. doi: 10.1109/iotm.001.2300181.

[8] H. Arif, A. Kumar, M. Fahad, and H. K. Hussain, "Future Horizons: AI-Enhanced Threat Detection in Cloud Environments: Unveiling Opportunities for Research," *International Journal of Multidisciplinary Sciences and Arts*, vol. 2, no. 2. Information Technology and Science (ITScience), pp. 242–251, Jan. 15, 2024. doi: 10.47709/ijmdsa.v2i2.3452.

[9] Y. Wang, H. Liu, Z. Li, Z. Su, and J. Li, "Combating Advanced Persistent Threats: Challenges and Solutions," *IEEE Network*, vol. 38, no. 6. Institute of Electrical and Electronics Engineers (IEEE), pp. 324–333, Nov. 2024. doi: 10.1109/mnet.2024.3389734.

[10] G. Giacomello and O. Preka, "Targeting Reputation: A New Vector for Attacks to Critical Infrastructures," *Computer and Information Science*, vol. 14, no. 3. Canadian Center of Science and Education, p. 63, Jul. 28, 2021. doi: 10.5539/cis.v14n3p63.

[11] I. Linkov *et al.*, "Measurable Resilience for Actionable Policy," *Environmental Science & Technology*. American Chemical Society (ACS), p. 130903081548008, Sep. 03, 2013. doi: 10.1021/es403443n.

[12] Md Rokibul Hasan, Md Sumon Gazi, and Nisha Gurung, "Explainable AI in Credit Card Fraud Detection: Interpretable Models and Transparent Decision-making for Enhanced Trust and Compliance in the USA," *Journal of Computer Science and Technology Studies*, vol. 6, no. 2. Al-Kindi Center for Research and Development, pp. 01–12, Apr. 06, 2024. doi: 10.32996/jcsts.2024.6.2.1.

[13] P. Kumar, D. Javeed, R. Kumar, and A. K. M. N. Islam, "Blockchain and explainable AI for enhanced decision making in cyber threat detection," *Software: Practice and Experience*, vol. 54, no. 8. Wiley, pp. 1337–1360, Feb. 19, 2024. doi: 10.1002/spe.3319.

[14] B. Dash, "Zero-Trust Architecture (ZTA): Designing an AI-Powered Cloud Security Framework for LLMs' Black Box Problems," *SSRN Electronic Journal*. Elsevier BV, 2024. doi: 10.2139/ssrn.4726625.

[15] K. Zisis, E. Pavi, M. Geitona, and K. Athanasakis, "Real-world data: a comprehensive literature review on the barriers, challenges, and opportunities associated with their inclusion in the health technology assessment process," *Journal of Pharmacy*

& Pharmaceutical Sciences, vol. 27. Frontiers Media SA, Feb. 28, 2024. doi: 10.3389/jpps.2024.12302.

[16] S. E. Galaitsi, J. M. Keisler, B. D. Trump, and I. Linkov, "The Need to Reconcile Concepts that Characterize Systems Facing Threats," *Risk Analysis*, vol. 41, no. 1. Wiley, pp. 3–15, Aug. 20, 2020. doi: 10.1111/risa.13577.

[17] M. L. Cummings, "A Taxonomy for AI Hazard Analysis," *Journal of Cognitive Engineering and Decision Making*, vol. 18, no. 4. SAGE Publications, pp. 327–332, Jan. 09, 2024. doi: 10.1177/15553434231224096.